DanceFix: Fixing Abnormal Challenges of Human Pose in Dance Assessment

Journal:	*Transactions on Visualization and Computer Graphics*
Manuscript ID	TVCG-2022-10-0554
Manuscript Type:	Regular
Keywords:	Action evaluation, anomaly correction, optical flow, part tracking, dance neatness evaluation

DanceFix: Fixing Abnormal Challenges of Human Pose in Dance Assessment

Xiao Ke, *Member, IEEE,* Huangbiao Xu, Peirong Xu, and Wenzhong Guo, *Member, IEEE*

Abstract—The objective and fair evaluation of performances and competitions is a common pursuit and challenge in human society, and the application of computer vision technology in real-world scenarios brings a ray of hope for this purpose. Nevertheless, it is still a challenging task to evaluate various types of behaviors accurately, such as performances and competitions, due to anomalies such as human occlusion and motion blur that hinder the process. To address these hindrances, our DanceFix proposes a new anomalous skeletal data correction method, called bidirectional spatial-temporal context optical flow correction (STCF), which exploits the consistency and complementarity of motion information between the two modalities of optical flow and skeletal data to extract pixel-level motion changes and correct anomalous data. In addition, the attributes of human body parts in motion are usually not uniform, we propose a part-level dance dataset (Dancer Parts) and part-level motion feature extraction based on task decoupling (PFTD), aiming to extract human limb-level motion information and improve the confidence of temporal information and accuracy of correction for abnormalities. Finally, we present the Dancing-Neatly-in-Virtual dataset (DNV), which simulates fully neat group dance scenarios and anomalous challenges to provide credible labels and validation methods for dance assessment. To the best of our knowledge, this is the first work to develop quantitative criteria for assessing dance neatness. Experimental results show that the proposed method can effectively correct anomalous skeletal points, flexibly embed and improve the accuracy of existing pose estimation algorithms, and fully validate the correction effect of the method on various dance scenes, DNV datasets, and video-based JHMDB datasets.

Index Terms—Action evaluation, anomaly correction, optical flow, part tracking, dance neatness evaluation.

1 INTRODUCTION

Human action recognition and evaluation is an important research area in computer vision, which has been widely used in real-world scenarios, such as human-computer interaction, video monitoring, and video retrieval. In recent years, the application of action recognition technology in sports and dance entertainment [1], [2] has slowly stepped into the limelight, and with the increased attractiveness of the game industry, games [3], [4] based on human interaction such as dance and sports have also extended the research in motion evaluation.

In this paper, we focus on the challenges involved in the problem of action recognition and assessment, taking the dance action assessment as an example, and proposing targeted solutions. Several modalities are available to study human action, such as RGB video [5], [6], optical flow [7], [8], shape [9], and skeletal data [10], [11], but these works study only one modality, ignoring consistency and complementarity of action information among the modalities. Compared with other modalities, the current commonly used skeletal data is lightweight structural data, which is not easily affected by the background and has higher computational efficiency and robustness. But the acquisition of skeletal data in practical applications depends on the detection accuracy of the pose estimation algorithms. The human dance movements studied in this paper are more variable compared with other common movements and have more complex motion information, which are more prone to abnormalities such as self-masking, external masking, and motion blurring leading to inaccurate skeletal data extraction (Fig. 1). This skeletal information with low credibility will reduce the accuracy of the dance neatness assessment. In addition, we find inconsistent motion properties of body parts during human motion when processing abnormal skeletal data. The human body has a flexible articulated structure, and each limb part is interconnected but often has different movement properties. It is difficult to achieve uniform changes in all body parts during movement. Such differences are even more obvious in dance movements. For example, there are more hand-related dance movements, which are usually more flexible and more intense, with complex movement change scenarios and large deformations; while the legs change more gently, with small postural deformations in certain sequences. If the different movement characteristics of each part are treated in the same way, it often leads to inaccurate detection of skeletal points that cause major disruption to the assessment work.

Besides, in the real world, dance movement evaluation is not only entertaining and interesting but also has important practical value. In performance and competition, people's evaluation of the same movement sequence is directly influenced by personal subjective consciousness, and even industry experts cannot give the same and convincing scores. While the method of using machines to evaluate movement sequences under fixed parameter criteria can fully ensure objectivity and fairness. Research in this area [12], [13] can be widely used in dance performances, sports competitions, troop exercises, and other scenarios where objective assessment is a better alternative to subjective

- *Xiao Ke, Huangbiao Xu, Peirong Xu, and Wenzhong Guo are with the Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing, College of Computer and Data Science, Fuzhou University, Fuzhou 350116, China, also with the Key Laboratory of Spatial Data Mining & Information Sharing, Ministry of Education, Fuzhou 350003, China (e-mail:kex@fzu.edu.cn; 211027092@fzu.edu.cn; 211027114@fzu.edu.cn; guowenzhong@fzu.edu.cn).*
- *Wenzhong Guo is the corresponding author.*

assessment, more importantly, it can provide an alternative perspective for thinking about research in the field of action recognition and assessment. However, there has been a lack of credible labeling data and unified quantitative standards in this area of research, which makes it difficult to achieve accurate quantitative assessment of movement sequences, preventing the effective dissemination of research results and concepts. For example, in this paper, most of the dance scenes that look neat at a glance are not neat after a closer look, and the dance scenes that are 100% neat in reality are difficult to find and do not even exist (Fig. 2), making it difficult to define the exact neatness of dance data.

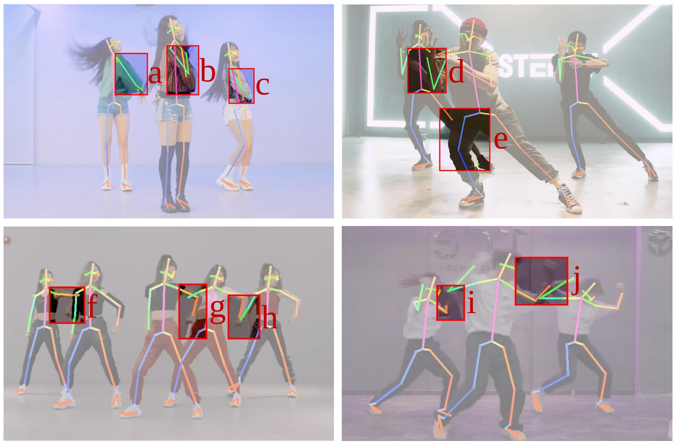


Fig. 1. Common anomalies in fixed-camera dance. Where a, b, c are self-masking, d, e, f, g, j are external-masking, h is limb-joining, i, j are motion blur.

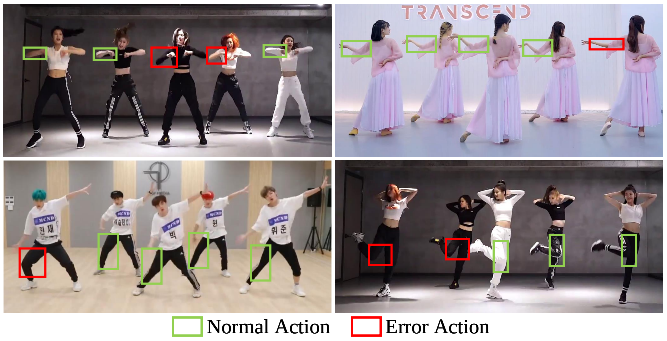


Fig. 2. Group dancing is difficult, if not impossible, to achieve completely neat. The green boxes are normal actions, the red boxes are errors.

Based on the above analysis, we summarize the main challenges for dance movement assessment as follows:

1) How to solve the shortcomings of a single modality and improve the accuracy of human motion information, especially the extraction of motion information under abnormal conditions such as occlusion and motion blur?
2) How to eliminate the influence of non-uniform motion properties of body parts during human movement and adaptively extract motion information from each part with various degrees of strenuousness?
3) How to obtain credible labeling data and quantify uniform evaluation criteria for group movements?

In this paper, we investigate dance movement assessment based on the three main challenges mentioned above. First, to address the problem of inaccurate extraction of current mainstream modality–skeletal data under anomalous conditions, we exploit the property that the motion information expressed among the modalities is consistent with each other. Among them, optical flow can acutely capture motion information on horizontal and vertical axes by detecting pixel intensity changes in continuous frames of images, and calculate the pixel changes before and after the appearance of anomalies, which has the potential to solve anomalies such as occlusion [14]. Therefore, we propose a new method for anomalous skeletal data correction called bidirectional spatial-temporal context optical flow correction (STCF). Our method first extracts the credible pre- and post-temporal context optical flow information of anomalous frames. Then calculates the spatial motion changes of skeletal point pixels with the optical flow information, using the optical flow to compensate for the defects of skeletal data, the most computationally efficient modality, to maximize the benefits. We can thus achieve the correction of anomalous skeletal data by combining the motion consistency of both modalities, skeletal data and optical flow.

Second, to extract the motion properties of flexible human body parts more effectively, we decouple the complex human body into limb parts based on the idea of task decoupling and conduct a study focusing on the “instance-level” information of human body parts. The detection and practice of “instance-level” body parts is already an important research area [15], and different levels of body part information are important for different problems [16], [17], [18]. Therefore, we construct a part-level dance movement dataset (Dancer Parts) for the difference in motion properties of body parts during human movement and propose a part-level motion feature extraction based on task decoupling (PFTD). Through PFTD, we extract the motion properties of human body parts to determine information such as candidate skeletal point regions and motion changes, and use this information to further detect the before and after frames of abnormal frames, which makes the spatial-temporal optical flow information of STCF more reliable and improves the credibility of correction and the accuracy of dance neatness assessment.

Third, to obtain a quantitative research standard and credible labeling data to enhance the reusability and feasibility of our work, we construct a dataset that simulates a fully neat virtual dance scene, called Dancing-Neatly-in-Virtual (DNV). In this way, quantitative criteria can be developed for dance assessment, facilitating research in the field and validating the accuracy and reliability of our DanceFix. We validate the feasibility of automatic quantitative dance neatness assessment and the validity of STCF and PFTD on the DNV, and further validate the effectiveness of the method in improving the accuracy of pose estimation on a publicly available dataset.

Generally, the main contributions of this paper are summarized as follows:

- We propose a new anomalous skeletal data correction method, called bidirectional spatial-temporal context optical flow correction (STCF), to combat anomalies such as occlusion and motion blur and improve the accuracy of skeletal data.

- To adapt to the variability of the motion of human body parts, we propose a part-level motion feature extraction based on task decoupling (PFTD) to extract information such as candidate regions of skeletal points and motion changes of human body parts to obtain more accurate spatial-temporal information. Further, to effectively learn the information at the human body part level, a corresponding part-level dance movement dataset (Dancer Parts) is constructed, which consists of 65 dance video clips with rich dance types, different dancers, and complex scenes, and the dancers' left hand, right hand, left leg, right leg, and torso are "instance-level" labeled. PFTD effectively improves the reliability of spatial-temporal context information and the correction accuracy of STCF.
- We construct a simulated fully neat dance dataset (DNV), which develops a unified criterion for automatically quantifying dance neatness assessment, enabling the assessment of movement neatness for group dances. To the best of our knowledge, this is the first work to develop quantitative criteria for assessing dance neatness. Unlike other dance datasets, DNV provides reliable data of 100% neatness and facilitates the construction of dance scenarios that simulate abnormal challenges, which can effectively validate the accuracy of neatness assessment methods.
- We conduct a comprehensive evaluation on the DNV dataset proposed in this paper to verify the effectiveness of our methods, and further, validate the effectiveness of our methods on the public dataset JHMDB for the correction of abnormal skeletal data. The experimental results show that DanceFix can effectively improve the detection accuracy of the existing pose estimation algorithms.

The rest of this paper is organized as follows. In Section 2, we review the related work, and in Section 3, we present the specific framework and detailed information of our DanceFix. In Section 4, we evaluate the validity of our method. Finally, we conclude in Section 5.

2 RELATED WORKS

2.1 Action Recognition

Due to the lightweight nature of skeletal data and posture estimation techniques that have matured in recent years, skeletal-based action recognition research has become a mainstream trend. The data-driven approach for human feature extraction using deep learning techniques gradually replaces the traditional manual feature approach [19], [20], and there are three main neural networks for skeleton-based action recognition: Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and Graph Convolutional Network (GCN). RNN-based approaches are usually modeled to capture temporal feature sequences between frames, e.g. Du et al. [21] and Wang et al. [22] build RNN architectures from the hierarchy of human physical structures and joint geometric relationships to capture spatial-temporal feature information, respectively. CNN-based methods [23], [24], on the other hand, encode and convert skeletal data into pseudo-images to exploit the excellent image information processing capability of CNNs. In contrast, in recent years, GCN-based methods [11], [25] have begun to explore the human skeletal architecture and refine the linkage relationships of bones and joints into graphical structures. Compared with the sequence information of RNNs and the pseudo-images of CNNs, the graphs utilized by GCNs are more consistent with the topological structure of the human body and better represent the dependencies between skeletal data.

Although skeleton-based action recognition techniques are becoming more and more mature, the quality of skeletal data still limits these methods in practical applications, and pose estimation methods often perform poorly in the face of scenarios such as occlusion and motion blur. To solve the anomaly problem of using a single skeletal data modality, our STCF captures motion information before and after anomalies using optical flow estimation and corrects anomalous skeletal data by using the consistency and complementarity of motion information between multiple modalities to obtain high-quality skeletal data for skeletal-based action recognition and evaluation studies.

2.2 Human Keypoint Detection

Human keypoint detection from images or videos, also known as pose estimation, has been an important task in computer vision. Since many large-scale pose estimation benchmarks contain only image information, such as COCO [26], MPII [27], etc., a large amount of pose estimation work [28], [29], [30] is performed mainly on single-frame images, and these methods can only estimate frame-by-frame when processing video and often perform poorly in the face of occlusion and motion blur scenes common to video. In recent years, more promising results have been obtained by modeling temporal features based on video-based pose estimation to learn and exploit the temporal information of videos. In particular, Song et al. [31] proposed structured models for end-to-end training of human poses in videos. Luo et al. [32] combined long short-term memory (LSTM) and convolutional pose machine (CPM) [33] for rewriting multi-stage CNN to RNN using shared weights to extract pose spatial-temporal features to improve video processing speed and pose estimation quality. Nie et al. [34] introduced a lightweight distiller to transfer pose information between continuous frames using temporal features of the previous frame to guide the next frame. Recently, more relevant to our study, Dang et al. [35] designed lightweight plug-ins to model joint relationships to learn joint correlations, combining temporal dynamics to transfer pose semantic features from non-occluded frames to occluded frames.

However, the utilization of temporal features by existing methods is often limited to one frame before and after to directly help localize the current frame. While we find that anomalies such as occlusion and motion blur often lead to inaccurate pose estimation in multiple continuous frames in human behavior, especially in complex motion scenes

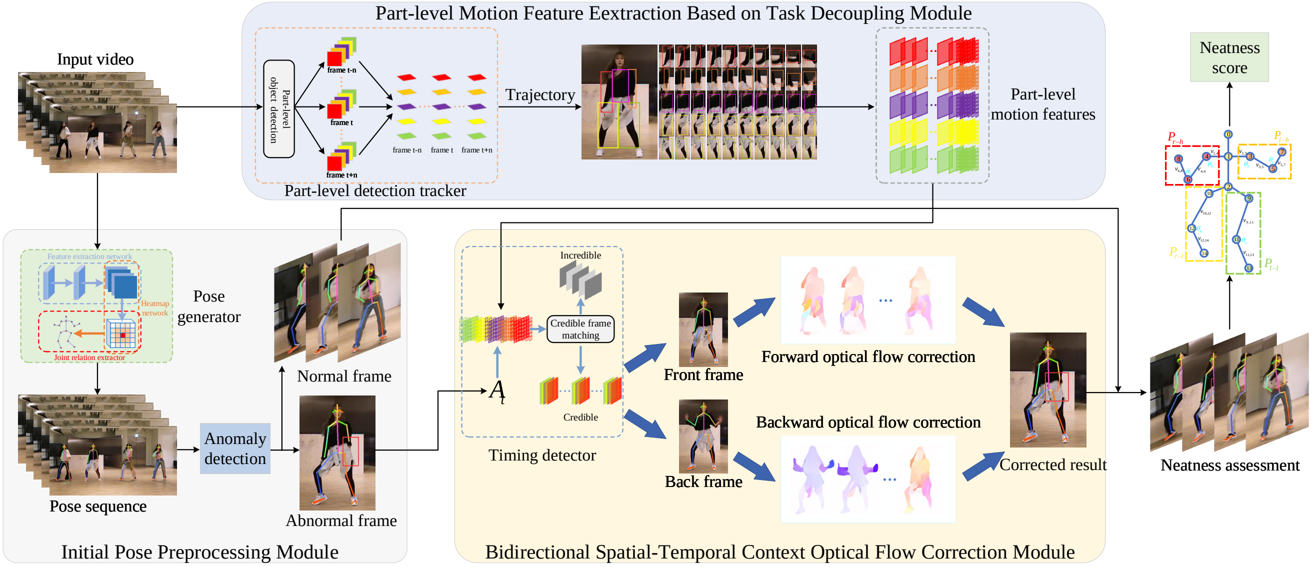


Fig. 3. Block diagram of DanceFix. The input video sequence is pre-processed to estimate the initial pose and perform anomalous skeletal data detection, and the PFTD module extracts part-level motion information to assist in the search for credible spatial-temporal context information. Then, the STCF module is used to correct the abnormal skeletal data based on the spatial-temporal context optical flow information, and finally to complete the dance movement neatness assessment.

such as dance and sports competition. For this reason, our method expands the field of view for extracting spatial-temporal features, excludes unreliable spatial-temporal features, and uses optical flow estimation to capture motion consistency to correct anomalous skeletal data.

2.3 Multiple Object Tracking

Multiple object tracking (MOT) has great market potential and academic value and has been widely interested by researchers. The classical MOT approach is based on target detection and data association (TBD), using a target detector to detect target bounding boxes frame by frame and a specific data association approach to identify the target identity. Bewley et al. [36] combined the Kalman filter and the Hungarian algorithm to propose a crude but effective MOT framework. Zhou et al. [37] used the detection results of previous frames to estimate the motion of the current frame and established tracking links for continuous frames based on the center position offsets. Zhang et al. [38] retain low-score detection frames to remove background interference and effectively track difficult targets such as occlusion and motion blur. On the other hand, some JDE framework methods [39], [40] use a unified model to link target detection and tracking to improve tracking efficiency. In addition, some recent methods [41], [42] try to apply the Transformer to the tracking task. The ever-advancing MOT techniques effectively create target trajectory tracking sequences that can provide information about human motion for tasks, such as action recognition and evaluation.

In this paper, we want to track dancers' motion trajectories to assist in correct abnormal skeletal data. However, all these MOT methods only track the whole body or head of the target, which is less applicable to scenarios with complex limb movements such as dance, sports. For this reason, we train a human part-level tracker based on the MOT task to track human parts at a finer level and capture the differences in the motion properties of each part.

3 METHOD

As shown in Fig. 3, our proposed dance neatness evaluation method DanceFix can be divided into three parts: initial pose preprocessing module, part-level motion feature extraction based on task decoupling module (PFTD), and bidirectional spatial-temporal context optical flow correction module (STCF). We first preprocess the video sequence for pose extraction to obtain the initial pose sequence and detect the abnormal skeletal data. We then input the detected abnormal frames into the spatial-temporal information detector to obtain the spatial-temporal context motion information on its pre- and post-temporal sequences. Meanwhile, we input the video sequence into PFTD to extract the limb-level motion information and fuse it with the spatial-temporal motion information to obtain more accurate skeletal candidate regions. Then, we obtain the optical flow motion information corresponding to the spatial-temporal context information. Finally, the optical flow information is used to correct within the skeletal candidate regions to obtain more accurate skeletal information for motion neatness assessment.

3.1 Bidirectional Spatial-Temporal Context Optical Flow Correction

In the research of skeletal-based action recognition, anomalies such as occlusion and motion blur lead to undetectable skeletal data and low detection accuracy, which greatly limits the upper limit of related research and practical applications. To improve the credibility of skeletal data, we propose bidirectional spatial-temporal context optical flow correction (STCF) for anomalous skeletal data.

For an input video sequence I_v, assume that it includes N frames, i.e. $I_v = \sum_{t=1}^{N} I_t, I_t \in {}^{H\times W\times 3}$. First, we generate the initial pose using a pre-trained human pose estimator for the input video sequence. Then, anomaly detection is performed on the initial pose. Specifically, we consider the low confidence or the large variance in the confidence sequence of the human body in the initial pose information, and the large instantaneous motion rate of the skeleton between two continuous frames as anomalous skeletal information that needs to be corrected. We summarize this process as follows:

$$A_t = \sum_k p_t^k = D(E(I_t)) \tag{1}$$

where A_t denotes the set of abnormal skeletal data in the detected abnormal frame I_t, p_t^k denotes the kth abnormal skeletal point in the video frame I_t, and $E(\cdot)$,$D(\cdot)$ are the human pose estimation and abnormality detection.

Since human motion has a certain degree of continuity and stability in a very short period, motion information within a time sequence in the vicinity of a frame is often related to the motion information of that frame, and many researchers have demonstrated the value of using temporal information [25], [35]. However, these researchers tend to utilize only the information of two continuous frames, while the current various keypoint detection algorithms often encounter anomalies in the detection of several continuous frames, even if only one frame appears to be anomalies. In practice, such continuous anomalous frames are more common since dance and sports are with more intense movements and complex changes in human behaviors. In this case, the skeletal information of the continuous before and after frames is often not credible, and the use of this information is not conducive to our anomaly correction. Therefore, for abnormal video frames, we extend our vision to detect anomalies frame by frame within a certain number of frames, exclude abnormal skeletal data until we find the nearest credible pre-sequence frame and post-sequence frame, and extract the past and future information of this motion.

Next, we extract the corresponding spatial-temporal context optical flow information between the credible before and after frames to the abnormal frames in order and reverse order, respectively. We then fuse the credible skeletal information of the before and after frames with the optical flow motion information, and extract their motion consistency. To improve the accuracy of the dance movement neatness assessment, we correct the skeletal point space information based on the pre-sequence frames and post-sequence frames and realize the recovery of the skeletal data of the current abnormal frame. Specifically, for the skeletal points p_i in the pre-sequence frames, we introduce the forward optical flow information $\overrightarrow{f}_i$ between this frame and the next frame, and use the motion consistency to calculate the skeletal points p_{i+1} in the next frame and use it as the input for the next calculation, and execute the calculation in order until the abnormal skeletal data is corrected. In this case, we start with the credible skeletal points p_{front} of the initial pre-process frame. In the other direction, we start from p_{back} to perform the inverse order correction on the motion information of post-sequence frames, and finally, fuse the order and reverse order correction results. In the case of extreme anomalies, we can only obtain a single credible pre-sequence or post-sequence information, and we make corrections based on this single information. The overall process is shown in Fig. 4, and we formalize the process as follows:

$$\overrightarrow{c_t} = \sum_{i=front}^{t-1} \overrightarrow{STCF}\left(p_i, \overrightarrow{f}_i, p_{i+1}\right) \tag{2}$$

$$\overleftarrow{c_t} = \sum_{i=back}^{t+1} \overleftarrow{STCF}\left(p_i, \overleftarrow{f}_i, p_{i-1}\right) \tag{3}$$

$$c_t = F\left(\overrightarrow{c_t}, \overleftarrow{c_t}\right) \tag{4}$$

where,$\overrightarrow{c_t}$ and $\overleftarrow{c_t}$ are the order and reverse order optical flow corrections for anomalous skeletal data p_t, respectively, $F(\cdot)$ is the correction fusion operations to obtain the final correction results c_t.

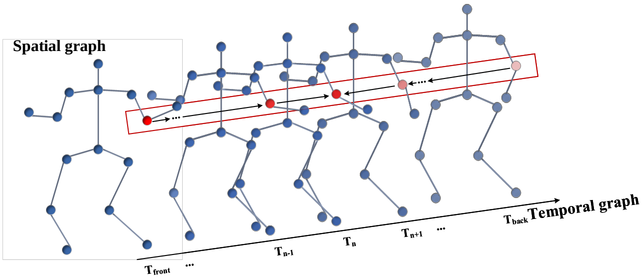


Fig. 4. Block diagram of bidirectional spatial-temporal context optical flow correction (STCF).

3.2 Part-level Motion Feature Extraction Based on Task Decoupling

The human body has a flexible articulated structure. Each body part can independently carry out different movements, which is the beauty of human movement. However, this property also leads to inconsistencies in the motion of the human body parts, with some parts having complex and fast movements and some parts having slow and calm movements. If all parts are treated in the same way, the differential information from local movements will be ignored. Therefore, we believe that we can make full use of this characteristic of each part and propose a part-level motion feature extraction based on task decoupling (PFTD). PFTD decouples the human body into individual parts based on the idea of task decoupling, and performs "instance-level" part detection and tracking of the human body to extract the body part-level motion information.

Although an "instance-level" human body part annotation dataset has been proposed in the research [15], it only focuses on the head, palms, and feet, and is not applicable to dance scenes with complex motion changes in various human body parts. To detect and track the dance movements in a more targeted way, we consider the common motion properties of human limbs based on dance movements. Since in many dance processes, dancers in groups often communicate with each other or with the audience in terms of eye contact and look conveys. The demand for neatness of the head is not high, so the head factor is excluded. Thus, we divide the human body into five parts: left and right hands, left and right legs, and torso. Although the dance neatness assessment algorithm introduced in Section 3.3 only calculates the neatness of each dancer's hands and legs, and the torso information is not directly involved in the assessment work. The self-masking and non-intentional alignment design of external masking in real dance scenes do not lead to the invisibility of the dancer's whole body, but present as the successive disappearance and reproduction of various body parts. The torso, as the center of the body, plays an important role in bridging various body parts and the re-identification of each part after the reappearance of the occlusion, and effectively improves the length and credibility of the tracking sequence of each part, so we also label the torso. After that, we build a part-level dance movement dataset (Dancer Parts) based on five limb parts, which includes a variety of dance types, scenes, variations, and other rich dance information (Fig. 5). Then, we train and validate our PFTD on this dataset, which first performs part-level detection for each dancer in the input video sequence, extracts the spatial information of each part, and tracks them. The obtained tracking sequence effectively preserves the spatial-temporal motion information of each part, which can provide candidate regions and motion changes reference information for several skeletal points involved in that part.

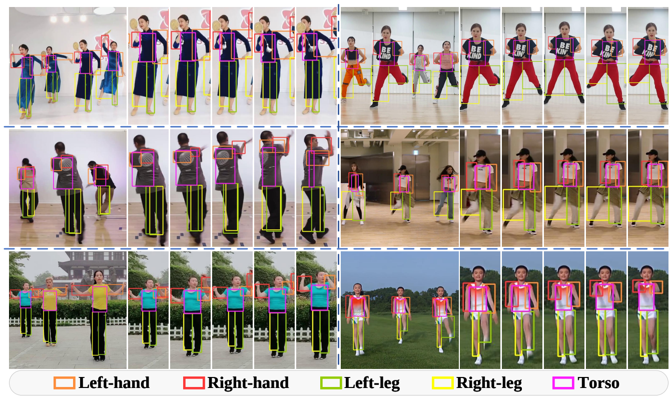


Fig. 5. An example of Dancer Parts dataset. We label the five parts of the body: left and right hands, left and right legs, and torso.

In addition, we find that the abnormal frame detection process of STCF still relies on the detection results of the keypoint detection algorithm, and the skeletal data in the temporal sequence near the abnormal frames are often unsatisfactory. Therefore, we use the part-level spatial-temporal tracking information obtained by PFTD to verify spatial location information and motion changes of the skeletal data of each limb of the dancer. We use the candidate regions obtained by PFTD to initially screen the skeletal points of the limb. And we further screen from them to find the skeletal points that are consistent with the overall motion change trend of the limb and match the relative movement of the articulated limb as credible spatial-temporal information, and the rest are regarded as incredible. The experiments prove that the incorporation of PFTD can eliminate the misjudgment caused by the incredible skeletal data, and make the correction effect of STCF sufficiently improved.

3.3 Dance neatness assessment

The modules proposed in this paper are ultimately dedicated to correcting abnormal skeletal data to achieve a valid and accurate neatness assessment of group dance, using a machine to provide an objective assessment of the movements. To this end, we propose a dance neatness assessment algorithm based on the extracted skeletal data. The algorithm comprehensively evaluates the motion, deformation process, and interconnection of human body parts during movement. And it acquires static features of each frame in the input video, deconstructs the human body into coarse-grained limb features and fine-grained joint features, and calculates the similarity of limb features based on cosine similarity and joint features based on a distance metric.

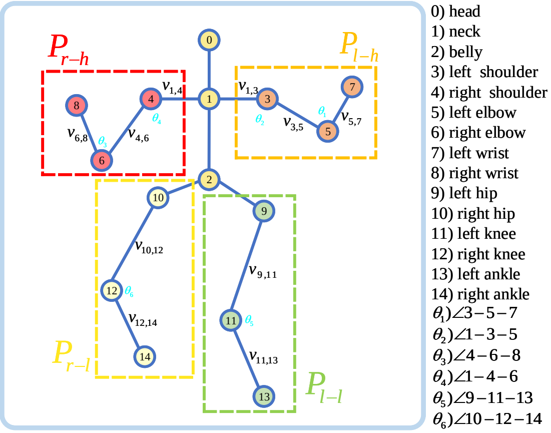


Fig. 6. An example of 2D poses of the human body numbering. We denote the human skeletal points as $P = \{0, 1, \cdots, 14\}$, each limb as vector $v_{p_{i_1},p_{i_2}}$, where p_{i_1},$p_{i_2} \in P$, each joint angle as $\theta = \{1, 2, \cdots, 6\}$, and split the body part into the left hand, right hand, left leg and right leg subsets, i.e. P_{l-h},P_{r-h},P_{l-l} and P_{r-l}.

As shown in Fig. 6, we number the 2D poses of the human body to describe the semantic features of each layer of the target. We first extract the limb-level features and split the human limb parts into eight limbs: left forearm, left upper arm, right forearm, right upper arm, left thigh, left calf, right thigh, and right calf. After that, we link the two skeletal points involved in each limb part to form the limb feature vector $v_{p_{i_1},p_{i_2}}$, where p_{i_1},p_{i_2} are the corresponding two skeletal points. The cosine similarity of the same limb feature vector among the dancers is calculated, and the higher the cosine similarity the closer the limbs are to parallel, and the more neatly the dancers move. This method effectively assesses movement neatness from limb-level characteristics, but each person has different degrees of limb coordination and each limb part has different degrees of ease of movement, with the thighs and upper arms easily achieving neatness, while the more flexible calf and forearms often have subtle differences. Considering only the limb-level characteristics tends to ignore these subtle differences and achieve a high degree of neatness, which is the challenge of neatness assessment and the dilemma that realistic group dance is difficult to achieve complete neatness. Therefore, we extract more fine-grained joint angle features, numbering the body joints commonly used in sports, and each joint angle is denoted by $\theta = \{1, 2, \cdots, 6\}$, representing left elbow angle, left shoulder angle, right elbow angle, right shoulder angle, left knee angle, and right knee angle, respectively. We normalize the joint angle values and use the L1 distance to calculate the degree of difference in the same joint angle feature among the dancers, measuring the characteristic distance: the smaller the distance, the neater the movement. Each joint angle is the angle between two limbs, and this method is based on the neatness of the relevant limb for a finer level of assessment, which amplifies the subtle differences and is more meaningful for the study of neatness assessment. With these two algorithms, the information from two grainy features of the human body is fused, and the correlation and neatness of limb and joint motion information are analyzed comprehensively. For a group dance with a number M, we formalize the neatness assessment in one frame as follows:

$$S_{\text{limbs}}(i_1, i_2) = \frac{1}{L}\sum_{l=1}^{L} \frac{v_l^{i_1} \cdot v_l^{i_2}}{\left|v_l^{i_1}\right| \times \left|v_l^{i_2}\right|} \tag{5}$$

$$S_{\text{joints}}(i_1, i_2) = \frac{1}{J}\sqrt{\sum_{j=1}^{J}\left(\frac{1}{1+\theta_j^{i_1}} - \frac{1}{1+\theta_j^{i_2}}\right)^2} \tag{6}$$

$$NS = \frac{1}{M}\sum_{i_1, i_2 \in M}\left(\lambda_1 S_{\text{limbs}}(i_1, i_2) + \lambda_2 S_{\text{joints}}(i_1, i_2)\right) \tag{7}$$

Where L and J are the number of limbs and joints, we set $L = 8$ and $J = 6$, i_1,i_2 are two different dancers, $v_l^{i_1}$ and $v_l^{i_2}$ are the lth limb vectors of the two dancers, $\theta_j^{i_1}$ and $\theta_j^{i_2}$ are the jth joint angles of the two dancers, λ_1,λ_2 are the weights of limb neatness S_{limbs} and joint neatness S_{joints}, respectively, and NS is the overall neatness score of the group.

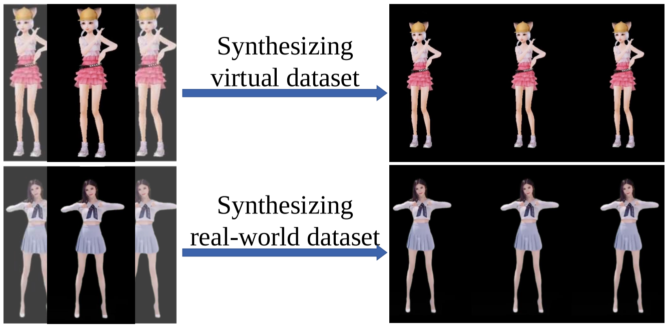


Fig. 7. Fully neat virtual dance movement dataset (DNV) generation method. Generate virtual and real-world datasets.

To validate the effectiveness of our neatness assessment method on credible data with quantitative metrics, we propose a DNV dataset based on the idea of motion synthesis [43], [44] to simulate a fully neat dance scenario. As shown in Fig. 7, we use individual virtual dance moves generated by, e.g., dance generation algorithms, games, software tools, etc., to be parsed and replicated, and reconstructed into virtual group dance moves, and further, we can also generate real group dances using real individual dances. In terms of both theoretical and practical validation, the new dance movement datasets generated are 100% neat, allowing for quantitative criteria of dance assessment, facilitating research in the field, and validating the accuracy and reliability of our neatness assessment method. To the best of our knowledge, this is the first work to develop quantitative criteria for dance neatness assessment.

4 EXPERIMENTS

In this section, we conduct extensive experiments on existing keypoint detection algorithms to verify the effectiveness of our proposed STCF and PFTD and perform comprehensive ablation experiments to explore the best methods and results for each part of the framework.

4.1 Datasets and evaluation metrics

We present our collected DNV dataset and Dancer Parts dataset in this section, where we validate the feasibility of our STCF and PFTD. In addition, we further validate our DanceFix on the publicly available video-based dataset JHMDB [45]. We show the content parameters of the three datasets in Table 1.

TABLE 1
Content parameters for DNV, Dancer Parts, and JHMDB datasets.

Name	FPS	Resolution	Clips	Min Length	Max Length	Total Length	Boxes
DNV	30	1280×720	50	60	300	5,300	238
Dancer Parts	30	1920×1080	65	103	1,562	13,666	1,190
JHMDB	15-40	320×240	928	15	40	31,838	928

4.1.1 *Dancing-Neatly-in-Virtual Dataset*

To define a uniform metric for neatness assessment and to validate the feasibility of our method in a fully neat environment, we extract dance information from common individual dance scenes, replicate and reconstruct them as group dances and collect them as a DNV dataset. The DNV dataset consists of the following: 50 group dance video sequences with different dance movements (25 virtual characters, 25 real characters), all dancers are fully visible and move neatly and consistently, with no background information, and a frame rate of 30 FPS.

The dance movements in this dataset conform to real dance common sense, and the movements are continuous and reproducible. The virtual individual dances come from dance games, avatar generation tools, etc. The real individual dances come from YouTube green screen dance videos and are reconstructed into group dances using video editing tools in unison. The virtual dancers conform to the body structure of real dancers and can extend the application scenarios of the dataset, which is meaningful for the study of application scenarios such as action recognition, keypoint detection, and other vision fields in games and virtual reality. To better represent the effect of anomalous skeletal data correction on a small base, we limited these video sequences to a small frame range with an average frame number of 70 frames. In addition, these group dances have multiple number sizes ranging from 2-7 people and no masking, which is to provide a complete 100% neat dataset to verify the feasibility of the neatness assessment method and the effectiveness of the correction method. Specifically, we also use random masking for this dataset to generate specific masking sub-datasets to simulate anomalous scenarios in dances, which facilitates the verification and improves the performance of DanceFix for anomalous cases.

4.1.2 Dancer Parts Dataset

To make full use of the limb part hierarchical motion information and solve the problem of non-uniform motion properties of each part, we collected a new dataset from YouTube, bilibili, and other video websites. The basic information of this dataset is as follows: 65 different group dance video sequences, 50 for the training set and 15 for the test set, labeled with the position information of detection boxes and the tracking sequence for each dancer's left hand, right hand, left leg, right leg, and torso, with a video resolution of 1920×1080 and a frame rate of 30 FPS.

The dataset follows the standard of the multiple object tracking dataset MOT20 [46] for labeling each part of the dancer, providing the spatial location information and video sequence tracking information of each part. To include more dance scenes for better generalization of our method, we intercept only a small segment of each dance video and ensure that it contains at least one continuous segment of normal dance movements that match the realistic dance scenes. To make the dataset balanced, we have male, female, mixed and children, adults, and elders by crowd categories, Chinese classical dance, Korean dance, European and American dance, radio gymnastics, square dance, and fitness gymnastics by dance type, normal unmasking, camera edge masking, dancer self-masking, external masking, and motion blur by abnormal scenes, and camera fixed dancer moving and dancer fixed camera moving by moving situation. The dataset has different costumes, lighting, and backgrounds, and contains rich information on dance types, balanced crowd categories, complex dance scenes, and sufficient abnormal information.

4.1.3 JHMDB

JHMDB [45] is one of the most commonly used datasets for video-based human pose estimation and contains a total of 928 video clips, labeling information for 15 skeletal points of the human body with 21 action categories. The dataset possesses a subset division (Sub-JHMDB) where all skeletal points of the whole body are visible in the video, with a total of 316 video clips. In addition, the dataset is divided into three different training and testing subsets, with a ratio of approximately 3:1 between the amount of training and testing. To facilitate comparison with previous work [33], [47], we evaluate our method in these three subset divisions and report the average results.

4.1.4 Evaluation indicators

To test the results of the dance neatness assessment and the effect of abnormal skeletal point correction in this paper, we use the neatness score (NS, introduced in Section 3.3) as our metric (i.e., the comprehensive neatness of each limb and joint between dancers). We validate the correction effect on the human body parts by calculating the NS of the left hand, right hand, left leg, and right leg using the limbs and joints involved in them, and calculate the overall NS of the whole body. In addition, we validate our migration ability for abnormal skeletal data correction on the JHMDB using the PCK [48] metric. For the evaluation of the information extraction effect of PFTD tracking sequences, we use the multiple object tracking accuracy (MOTA), the number of identity switches (IDS), identification F1 score (IDF1), and other related metrics, which are common and important in the field of MOT and can effectively evaluate the tracking performance.

4.2 Ablation experiments

In this section, we perform extensive experiments to evaluate the performance of the proposed STCF and PFTD and study the ablation experiments for different parts of the module chosen to obtain the best results and confirm the final overall framework. We perform the ablation experiments using AlphaPose [28] as the keypoint detection backbone. Although AlphaPose is an early pose estimation algorithm, its still competitive detection accuracy and the advantages of stable operation and high open-source quality make it has been widely used in application scenarios related to posing estimation. On the experimental side, correcting a non-optimal initial pose sequence helps DanceFix to better investigate the challenges in dance scenarios and improve the performance of the method.

4.2.1 Ablation experiment of STCF module

We conduct experiments on 10 video sequences of real human scenes randomly selected from the DNV dataset and average the results. We compare the STCF module with the common methods of correcting skeletal data. As shown in the Real-world Dataset of Table 2, compared with the methods of directly replacing the abnormal frame skeletal data with the credible temporal frames (denoted as *Replace*) and calculating the abnormal frame skeletal data based on the movement speed of the skeletal points in the temporal frames (denoted as *Line Speed*), our STCF is more effective in correcting the abnormal skeletal data of the dancer parts, achieving an NS increasing by 8.38% of the whole body compared with the original skeletal data of AlphaPose. This shows that the optical flow motion information has motion consistency with the skeletal data, and using the complementarity of the two modalities is beneficial to correct the abnormal skeletal data and improve the pose estimation accuracy.

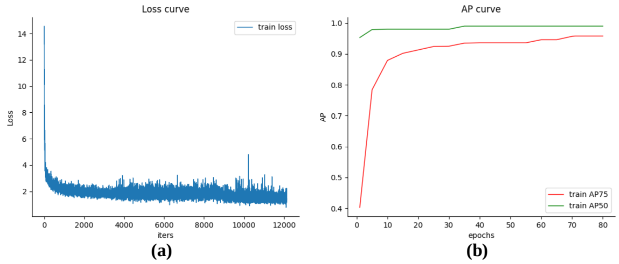
4.2.2 Ablation experiments of the PFTD module

The PFTD module trains a part-level tracker in our proposed Dancer Parts dataset, and we show its Loss and AP curve on the train set in Fig. 8 and its performance metrics on the test set in Table 3. Important metrics such as MOTA, IDS, and IDF1 achieve state-of-the-art results in the MOT domain, which indicates that our PFTD can effectively extract dancers' limb-level motion tracking sequences in dance scenes. To investigate the effect of the incorporation of the PFTD module on the correction results, we conduct experiments with the PFTD module for all three correction methods, denoted as *Replace+PFTD*, *Line Speed+PFTD*, and *STCF+PFTD*, respectively. The experimental results on Real-world Dataset in Table 2 show that the incorporation of the PFTD module improves the whole-body NS of the three methods by 1.09%, 0.34%, and 1.06%, respectively, which indicates that the motion tracking information of each body

TABLE 2
Ablation experiments with STCF and PFTD modules for the NS(%)↑ metric on DNV's Real-world Dataset and Virtual Dataset.

	Real-world Dataset					Virtual Dataset				
Methods	Left-hand	Right-hand	Left-leg	Right-leg	All-body	Left-hand	Right-hand	Left-leg	Right-leg	All-body
AlphaPose [28]	95.34	84.41	89.59	96.50	85.74	94.94	93.82	99.36	98.33	94.39
Replace	95.04	90.39	95.36	98.81	91.48	95.43	95.07	99.47	98.41	94.97
Replace + PFTD	**97.74**	89.17	95.36	98.82	92.57	95.53	**95.11**	**99.48**	98.41	95.00
Line Speed	95.15	90.09	95.36	98.82	91.88	95.44	94.90	99.44	98.41	94.89
Line Speed + PFTD	97.40	89.53	**95.37**	98.82	92.22	95.51	94.95	99.45	98.41	94.92
Ours(STCF)	95.03	92.22	95.36	**98.83**	93.19	95.66	94.99	**99.48**	98.40	**95.02**
Ours(STCF+PFTD)	97.71	**92.80**	95.36	**98.83**	**94.25**	**95.69**	95.00	**99.48**	**98.42**	**95.02**

part extracted by PFTD facilitates the acquisition of more credible before-and-after temporal frames, and that hierarchical motion information can identify credible skeletal point candidate regions. In addition, the NS of each limb part of the original skeletal data has obvious differences, which is a challenge brought by the non-uniform motion properties of each body part. Our method extracts the hierarchical motion features of each part, matches the part motion trajectories, and targets local corrections, which improve the NS of the left hand, right hand, left leg, and right leg by 2.37%, 8.39%, 5.77%, and 2.33%, respectively. The comprehensive correction of whole-body skeletal point data improved by 8.51% over the original data, which proved the effectiveness of our PFTD.



(a) (b)

Fig. 8. The Loss curve(a) and AP curve(b) of the part-level tracker on the Dancer Parts train set.

TABLE 3
The part-level tracker performance on the Dancer Parts test set.

Test data	Objects	MOTA↑	IDF1↑	IDs↓	MT↑	ML↓	FP↓	FN↓
Dancer Parts-51	4,375	86.2	83.4	29	96.0	0.0	269	302
Dancer Parts-52	2,950	76.9	83.4	7	68.0	0.0	300	375
Dancer Parts-53	1,090	64.0	76.9	11	60.0	0.0	147	234
Dancer Parts-54	1,200	76.6	82.0	4	60.0	0.0	82	195
Dancer Parts-55	3,270	82.9	85.6	13	66.7	0.0	232	314
Dancer Parts-56	3,900	93.4	94.0	12	100.0	0.0	101	139
Dancer Parts-57	3,400	92.2	92.8	6	100.0	0.0	111	148
Dancer Parts-58	2,620	58.2	67.7	50	60.0	0.0	350	694
Dancer Parts-59	4,875	83.2	78.1	33	88.0	0.0	374	415
Dancer Parts-60	2,640	73.7	80.2	17	60.0	0.0	253	426
Dancer Parts-61	1,230	83.7	90.0	2	80.0	0.0	68	130
Dancer Parts-62	3,690	71.2	48.2	50	60.0	0.0	370	644
Dancer Parts-63	1,890	64.9	63.8	48	60.0	0.0	240	378
Dancer Parts-64	3,645	68.9	48.2	43	60.0	0.0	420	470
Dancer Parts-65	4,525	68.4	69.8	32	60.0	0.0	551	846
OVERALL	45,300	77.6	75.8	357	73.8	0.0	4,021	6,241

4.2.3 Ablation experiments for a completely neat virtual dance dataset

To explore the connection and difference between reality and virtual, we conduct ablation experiments on two parts of DNV datasets built based on virtual and real characters. In Table 2, we can see that each method works better on the real character dataset than on the virtual dataset. The possible reason is that both the keypoint detection algorithm and PFTD are trained on the real person dataset, which are more accurate in detecting key points for real people and provides more credible temporal skeletal information. On the other hand, DanceFix also achieves good results on virtual datasets and demonstrates the feasibility of integration of this field with scenes such as VR and dance sports games, which also shows that our method is equally effective in virtual scenes.

4.2.4 Ablation experiments of neatness assessment methods

As shown in Table 4, we investigate whether the neatness assessment method used in this paper is reasonable. We conduct experiments using the full DanceFix (i.e., *STCF+PFTD*) on a fully neat real dance dataset, calculating cosine similarity for limb vectors and distance differences for joint angles. The experimental results show that limb neatness tends to present higher than joint neatness, probably because some body parts (e.g., upper arms, thighs) can easily reach neatness and make the cosine similarity reach higher values, while the differences in the remaining parts have more influence on the joint angle differences than the limb cosine similarity. It is a challenge to uncover and evaluate the subtle movement differences between dancers, and combining the results of both limb and joint neatness can provide a more valuable result for evaluation.

TABLE 4
Ablation experiments with different dance movement neatness assessment algorithms on the NS(%)↑ metric on the DNV dataset.

Methods	Left-hand	Right-hand	Left-leg	Right-leg	All-body
Limb neatness	97.86	94.28	99.96	99.97	98.02
Joint neatness	97.56	91.32	90.77	97.69	90.48
Avg neatness	97.71	92.80	95.36	98.83	94.25

4.2.5 Ablation experiment of correction range

In this subsection, we explore the effect of different ranges of abnormal skeletal point definitions on the correction effect. We denote the existence of high variance for the skeletal points confidence sequence of the dancer as *HV* and the low confidence of the skeletal points as *LC*. *HV+LC* means that dancers with high variance sequences correct only the skeletal points with low confidence. *All Points* means that all the skeletal points of the dancer are corrected for each frame, and PFTD means that all the skeletal points that do not match the part-level motion properties are corrected. The results in Table 5 show that low-confidence skeletal points tend to be anomalous, and *LC* improves the whole-body NS by 3.06% over *LV+LC*. In addition, the whole-body NS of *All Points* is further improved, which indicates that the skeletal points obtained based on the keypoint detection algorithm cannot rely solely on their confidence to determine whether they are abnormal or not, and there are often more abnormal skeletal points judged to be of high confidence. For this reason, we use the PFTD module to decouple the human motion information by each part, match all skeletal points with part-level motion properties, correct the skeletal points that failed to be matched, and finally achieve the best correction effect.

TABLE 5
Ablation experiments for different ranges of corrections and ranges of temporal information on the NS(%)↑ metric on the DNV dataset. Rows and columns represent different ranges of corrections and ranges of temporal information, respectively.

Methods	**4 frames**	**6 frames**	**8 frames**	**10 frames**	**All frames**
LV	87.09	87.10	87.12	87.13	86.95
LC	88.94	88.26	87.95	88.40	89.83
LV + LC	86.38	87.11	87.15	86.93	86.77
All Points	92.43	92.54	92.76	92.67	91.51
PFTD	92.45	93.48	92.94	93.49	**94.25**

4.2.6 Ablation experiments of the range of temporal information

We conduct ablation experiments on the credible range of temporal information, using 4 frames, 6 frames, 8 frames, 10 frames, and all frames as the credible range for correction. The results are recorded in Table 5. We can see that it is difficult to identify a uniform range of temporal information that can be utilized for different correction ranges, but after using our PFTD to filter the temporal frames matching the part-level motion properties, the confidence of temporal frames is significantly improved, and the best correction results are achieved using the full range of frames.

4.2.7 Ablation experiments of different optical flow estimation backbones

Optical flow motion information is one of the most important pieces of information in this paper, and to investigate the effect of the optical flow network on the final correction performance, we conduct ablation experiments on five different optical flow estimation backbones, FlowFormer [49], GMA [50], GMFlowNet [51], RAFT [52], and FlowNet2 [53], and the experimental results are given in Table 6. FlowFormer achieves the best results when performing optical flow estimation in the group dance scenario in this paper.

4.3 Results and comparison

4.3.1 Correction effects on state-of-the-art methods

We report the correct results after applying our method to state-of-the-art human pose estimation methods. On the same representative device environment (a Tesla V100, PyTorch 1.8.0), we use the open source code and pre-trained models of these state-of-the-art methods to obtain their pose detection results and obtain the initial dance movement neatness based on the pose sequence. We then use our method to detect and correct the initial skeletal sequence for abnormal skeletal data and show the experimental results in Table 7. The table has several blocks, and we list the initial dance movement neatness of all algorithms in the first row of each block, and the second row shows the correction effect of the algorithm after adding our method. We can see that our method has a good correction effect on each of the algorithms, with an average increase of 3.34% in whole-body NS. It is worth noting that there are cases of mediocre corrections for single body parts, possibly because each pose estimation algorithm has different detection effects on various parts of the body, and some of the initial pose sequences provide part-level skeletal data that are poorly detected.

TABLE 6
Ablation experiments with different optical flow estimation backbones on the NS(%)↑ metric on the DNV dataset.

Methods	Left-hand	Right-hand	Left-leg	Right-leg	All-body
FlowFormer	**97.71**	**92.80**	**95.36**	98.83	**94.25**
GMA	97.23	92.68	**95.36**	98.83	93.85
GMFlowNet	96.57	89.53	94.18	**99.96**	93.76
RAFT	97.32	92.57	**95.36**	98.82	93.89
FlowNet2	96.59	91.35	94.22	98.82	91.60

TABLE 7
Comparison of the correction effect of applying DanceFix to the latest pose estimation algorithm on the NS(%)↑ metric.

Methods	Left-hand	Right-hand	Left-leg	Right-leg	All-body
AlphaPose [28]	95.34	84.41	89.59	96.50	85.74
+DanceFix(Ours)	**97.71**	**92.80**	**95.36**	**98.83**	**94.25**
Hrnet [54]	**94.94**	93.45	91.65	92.90	89.56
+DanceFix(Ours)	94.33	**94.02**	**93.93**	**95.18**	**91.77**
UDP [55]	94.92	96.86	**94.23**	**97.69**	90.58
+DanceFix(Ours)	**96.88**	**97.51**	**94.23**	**97.69**	**92.24**
DEKR [56]	97.13	**93.10**	87.31	95.38	88.70
+DanceFix(Ours)	**97.57**	93.01	**93.07**	**96.53**	**91.30**
LiteHrnet [29]	**97.79**	95.90	94.22	95.38	92.90
+DanceFix(Ours)	97.68	**96.93**	**96.53**	**97.68**	**96.26**
DarkPose [30]	**97.16**	95.50	**96.54**	**100.00**	91.39
+DanceFix(Ours)	96.92	**96.16**	**96.54**	**100.00**	**92.39**
PVT [57]	93.92	84.71	93.06	99.98	87.41
+DanceFix(Ours)	**96.14**	**88.71**	**96.52**	**99.99**	**94.50**
PVT2 [58]	**94.54**	**96.77**	95.36	99.98	92.58
+DanceFix(Ours)	94.42	96.41	**96.52**	**99.99**	**94.37**
RSN [59]	95.35	93.17	89.60	97.67	87.36
+DanceFix(Ours)	**96.08**	**94.13**	**95.37**	**98.83**	**91.64**
Scnet [60]	**97.37**	95.47	94.22	96.53	91.71
+DanceFix(Ours)	**97.37**	**95.58**	**95.38**	**97.68**	**92.61**

4.3.2 Effectiveness comparison on public datasets

We further validate the effectiveness and migration capabilities of our methods on the video-based JHMDB dataset. We perform a modification based on SimpleBaseline2D (SBL) [47] and CPM [33]. For an accurate and efficient comparison, we perform initial pose estimation using pre-trained models based on three split subsets of these methods. We evaluate the impact of different detectors based on a comparison of Faster R-CNN [61] and Yolov3 [62]. In addition, we find that the PCK normalized by person size seems to be saturated, and using the classical PCK normalized by torso size [48] can better demonstrate the effectiveness of the method. We took a threshold value of 0.2 for reporting, i.e., PCK@0.2. As shown in Table 8, our method achieves good corrections with both different target detectors, and STCF improved on average by 2.60% after correction on SBL and 1.62% on average compared to CPM. And with the addition of the PFTD module (STCF+PFTD), we achieved an average improvement of 5.69% and 8.47% on SBL and CPM, respectively.

TABLE 8
Comparison of the correction effect of applying DanceFix to SBL and CPM on the JHMDB dataset on the PCK@0.2 metric.

Methods	PCK@0.2↑			
	Sub1	Sub2	Sub3	Avg
Detector:Faster R-CNN [61]				
SBL [47]	74.74	67.14	75.70	72.53
+Ours(STCF)	77.09(↑2.35)	70.09(↑2.95)	78.24(↑2.54)	75.14(↑2.61)
+Ours(STCF+PFTD)	**79.06(↑4.32)**	**74.06(↑6.92)**	**80.67(↑4.97)**	**77.93(↑5.40)**
CPM [33]	59.19	52.78	61.67	57.88
+Ours(STCF)	61.09(↑1.90)	54.35(↑1.57)	63.29(↑1.62)	59.58(↑1.70)
+Ours(STCF+PFTD)	**67.71(↑8.52)**	**61.83(↑9.05)**	**69.54(↑7.87)**	**66.36(↑8.48)**
Detector:Yolov3 [42]				
SBL [47]	74.84	65.62	74.53	71.66
+Ours(STCF)	74.42(↑2.58)	68.50(↑2.88)	76.80(↑2.27)	74.24(↑2.58)
+Ours(STCF+PFTD)	**79.80(↑4.96)**	**73.22(↑7.70)**	**79.79(↑5.26)**	**77.64(↑5.97)**
CPM [33]	59.03	52.67	61.04	57.58
+Ours(STCF)	60.72(↑1.69)	54.4(↑1.37)	62.58(↑1.54)	59.11(↑1.53)
+Ours(STCF+PFTD)	**67.83(↑8.80)**	**61.51(↑8.87)**	**68.73(↑7.69)**	**66.03(↑8.45)**

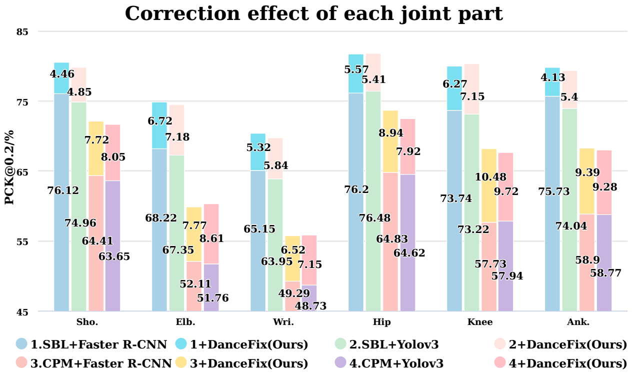


Fig. 9. Correction results of the DanceFix on the JHMDB test set of each joint part. The bottom of each bar shows the initial effect of the original method, while the top shows the increased effect after adding ours.

In addition, we report the average correction effect across joints on the JHMDB in Fig. 9. Our method achieves a more balanced correction effect across joints, with the wrist tending to be less effective, possibly because the wrist region is more flexible and more susceptible to motion blur.

4.3.3 Visualizations of correction effect

We show the typical anomaly correction effect of DanceFix in the dance scene in Fig. 10 and Fig. 11, and the typical anomaly correction effect of DanceFix on the JHMDB dataset in Fig. 12. We report our correction effect for abnormal cases, such as occlusion, motion blur, and environmental influence of the scene. From the visual effects, we can see that DanceFix can effectively solve the abnormal skeletal data in the human pose estimation scene. We show more visualizations of correction effect in the Appendix.

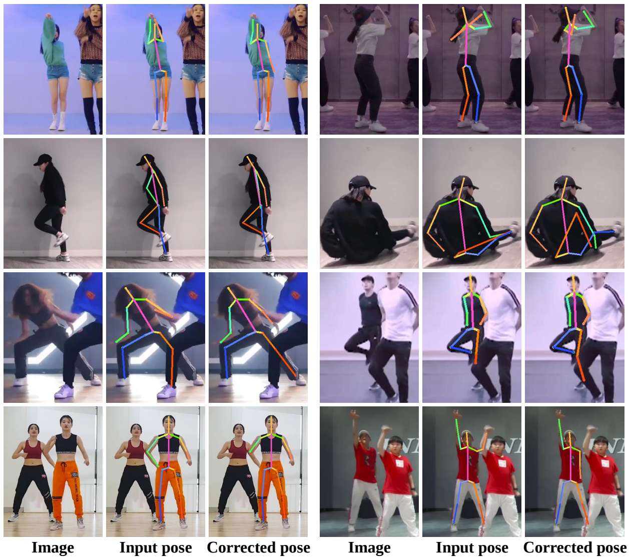


Fig. 10. Correction results of the DanceFix in the dance scene regarding the anomalies of self-masking (rows 1-2), external masking (row 3) and limb-joining (row 4).

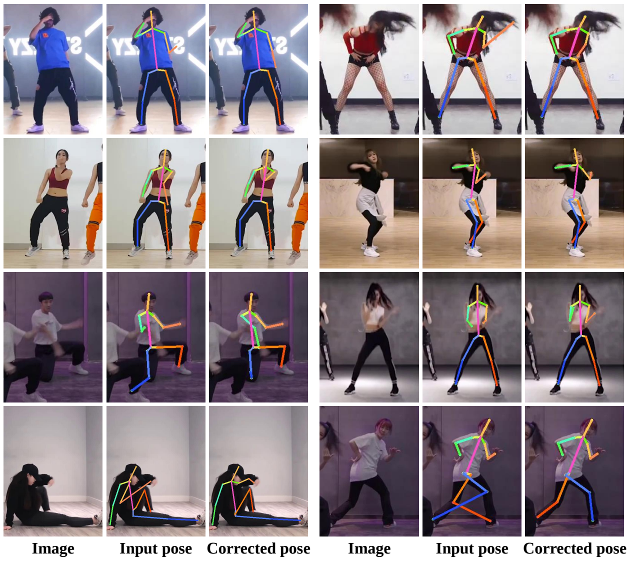


Fig. 11. Correction results of the DanceFix in the dance scene regarding environmental factors such as hair, costume and background (rows 1-2), motion blur (row 3) and common errors (row 4).

5 Conclusion

In this paper, we focus on exploring the challenges faced in action recognition and evaluation using dance movement assessment as an example, aiming to correct inaccurate skeletal data due to anomalies such as occlusion and motion blur. To address this important issue, we first propose a bidirectional spatial-temporal context optical flow correction module (STCF), which is a new method for correcting anomalous skeletal data by exploiting the motion consistency and complementarity of the two modalities of optical flow and skeletal data. To obtain credible spatial-temporal information, we collect the Dancer Parts dataset and train a part-level dance movement tracker based on this dataset,

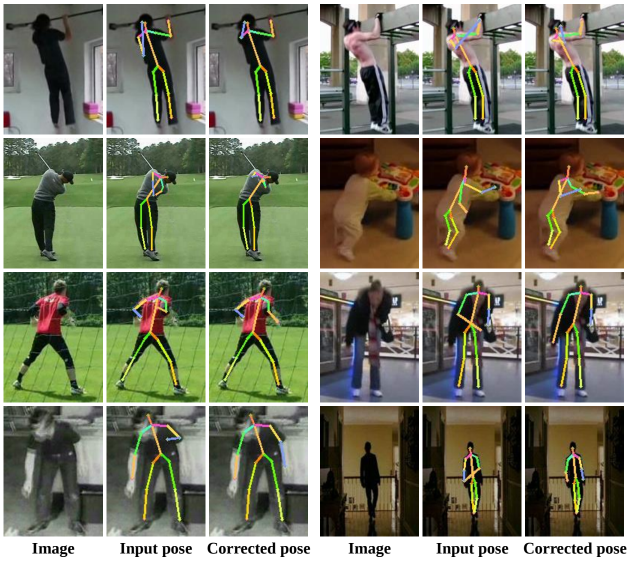


Fig. 12. Correction results of the DanceFix on the JHMDB test set regarding anomalies such as occlusion (rows 1-2), motion blur (row 3), low light and environmental influence (row 4).

which constitutes a part-level motion feature extraction based on task decoupling module (PFTD) to extract the body part-level motion information to obtain more credible temporal frames. After that, we collect the DNV dataset to define a unified criterion for automatic quantitative dance movement assessment to examine the neatness assessment and correction effects. We conduct extensive experiments on the proposed modules to find the best correction effect, and the experimental results show that our DanceFix can flexibly plug into the latest pose estimation methods and all of them can efficiently correct abnormal skeletal data and improve the accuracy. Further, to the best of our knowledge, we are the first to define automatic quantitative criteria for dance movement assessment, and we hope that our work will provide a meaningful and interesting idea for the movement assessment of dance and sports competitions with a highly realistic value.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61972097 and U21A20472, in part by the National Key Research and Development Plan of China under Grant 2021YFB3600503, in part by the Natural Science Foundation of Fujian Province under Grant 2021J01612 and 2020J01494, in part by the Major Science and Technology Project of Fujian Province under Grant 2021HZ022007, in part by the Industry-Academy Cooperation Project of Fujian Province under Grant 2018H6010, in part by the Fujian Collaborative Innovation Center for Big Data Application in Governments, and in part by the Fujian Engineering Research Center of Big Data Analysis and Processing. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers.

References

[1] X. Guo, Y. Zhao, and J. Li, “Danceit: music-inspired dancing video synthesis,” *IEEE Transactions on Image Processing*, vol. 30, pp. 5559–5572, 2021.

[2] Y. Wu, J. Lan, X. Shu, C. Ji, K. Zhao, J. Wang, and H. Zhang, “ittvis: Interactive visualization of table tennis data,” *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 709–718, 2017.

[3] E. Wu, M. Piekenbrock, T. Nakumura, and H. Koike, “Spinpong-virtual reality table tennis skill acquisition using visual, haptic and temporal cues,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 5, pp. 2566–2576, 2021.

[4] Z. Liu, L. Zhou, H. Leung, and H. P. Shum, “Kinect posture reconstruction based on a local mixture of gaussian process models,” *IEEE transactions on visualization and computer graphics*, vol. 22, no. 11, pp. 2437–2450, 2015.

[5] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[6] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[7] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” *Advances in neural information processing systems*, vol. 27, 2014.

[8] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933–1941.

[9] Y. Li, R. Xia, and X. Liu, “Learning shape and motion representations for view invariant skeleton-based action recognition,” *Pattern Recognition*, vol. 103, p. 107293, 2020.

[10] Z. Yang, Y. Li, J. Yang, and J. Luo, “Action recognition with spatio–temporal visual attention on skeleton image sequences,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2405–2415, 2018.

[11] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, “An attention enhanced graph convolutional lstm network for skeleton-based action recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1227–1236.

[12] X. Bruce, Y. Liu, K. C. Chan, Q. Yang, and X. Wang, “Skeleton-based human action evaluation using graph convolutional network for monitoring alzheimer’s progression,” *Pattern Recognition*, vol. 119, p. 108095, 2021.

[13] T. Polk, J. Yang, Y. Hu, and Y. Zhao, “Tennivis: Visualization for tennis match analysis,” *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 2339–2348, 2014.

[14] D. Poux, B. Allaert, N. Ihaddadene, I. M. Bilasco, C. Djeraba, and M. Bennamoun, “Dynamic facial expression recognition under partial occlusion with optical flow reconstruction,” *IEEE Transactions on Image Processing*, vol. 31, pp. 446–457, 2021.

[15] L. Yang, Q. Song, Z. Wang, M. Hu, and C. Liu, “Hier r-cnn: Instance-level human parts detection and a new benchmark,” *IEEE Transactions on Image Processing*, vol. 30, pp. 39–54, 2020.

[16] X. Wu, C. Li, S.-M. Hu, and Y.-W. Tai, “Hierarchical generation of human pose with part-based layer representation,” *IEEE Transactions on Image Processing*, vol. 30, pp. 7856–7866, 2021.

[17] Y. Wang, G. Li, H. Zhang, X. Zou, Y. Liu, and Y. Nie, “Panoman: Sparse localized components–based model for full human motions,” *ACM Transactions on Graphics (TOG)*, vol. 40, no. 2, pp. 1–17, 2021.

[18] P. Li, K. Aberman, R. Hanocka, L. Liu, O. Sorkine-Hornung, and B. Chen, “Learning skeletal articulations with neural blend shapes,” *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–15, 2021.

[19] R. Vemulapalli, F. Arrate, and R. Chellappa, “Human action recognition by representing 3d skeletons as points in a lie group,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 588–595.

[20] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*, vol. 1. Ieee, 2005, pp. 886–893.

[21] H. Wang and L. Wang, "Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4382–4394, 2018.

[22] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[23] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3288–3297.

[24] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, pp. 346–362, 2017.

[25] X. Hao, J. Li, Y. Guo, T. Jiang, and M. Yu, "Hypergraph neural network for skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 2263–2275, 2021.

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[27] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 2014, pp. 3686–3693.

[28] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2334–2343.

[29] C. Yu, B. Xiao, C. Gao, L. Yuan, L. Zhang, N. Sang, and J. Wang, "Lite-hrnet: A lightweight high-resolution network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 440–10 450.

[30] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7093–7102.

[31] J. Song, L. Wang, L. Van Gool, and O. Hilliges, "Thin-slicing network: A deep structured model for pose estimation in videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4220–4229.

[32] Y. Luo, J. Ren, Z. Wang, W. Sun, J. Pan, J. Liu, J. Pang, and L. Lin, "Lstm pose machines," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5207–5215.

[33] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[34] X. Nie, Y. Li, L. Luo, N. Zhang, and J. Feng, "Dynamic kernel distillation for efficient pose estimation in videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6942–6950.

[35] Y. Dang, J. Yin, and S. Zhang, "Relation-based associative joint location for human pose estimation in videos," *IEEE Transactions on Image Processing*, 2022.

[36] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3464–3468.

[37] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *European Conference on Computer Vision*. Springer, 2020, pp. 474–490.

[38] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," *ECCV*, 2022.

[39] W. Li, Y. Xiong, S. Yang, M. Xu, Y. Wang, and W. Xia, "Semi-tcl: Semi-supervised track contrastive representation learning," *arXiv preprint arXiv:2107.02396*, 2021.

[40] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "Fairmot: On the fairness of detection and re-identification in multiple object tracking," *International Journal of Computer Vision*, vol. 129, no. 11, pp. 3069–3087, 2021.

[41] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "Trackformer: Multi-object tracking with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8844–8854.

[42] P. Sun, J. Cao, Y. Jiang, R. Zhang, E. Xie, Z. Yuan, C. Wang, and P. Luo, "Transtrack: Multiple object tracking with transformer," *arXiv preprint arXiv:2012.15460*, 2020.

[43] T. Yin, L. Hoyet, M. Christie, M.-P. Cani, and J. Pettre, "The one-man-crowd: Single user generation of crowd motions using virtual reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 5, pp. 2245–2255, 2022.

[44] Z. Wang, J. Chai, and S. Xia, "Combining recurrent neural networks and adversarial training for human motion synthesis and control," *IEEE transactions on visualization and computer graphics*, vol. 27, no. 1, pp. 14–28, 2019.

[45] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3192–3199.

[46] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé, "Mot20: A benchmark for multi object tracking in crowded scenes," *arXiv preprint arXiv:2003.09003*, 2020.

[47] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 466–481.

[48] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2878–2890, 2012.

[49] Z. Huang, X. Shi, C. Zhang, Q. Wang, K. C. Cheung, H. Qin, J. Dai, and H. Li, "FlowFormer: A transformer architecture for optical flow," *ECCV*, 2022.

[50] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. Hartley, "Learning to estimate hidden motions with global motion aggregation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9772–9781.

[51] S. Zhao, L. Zhao, Z. Zhang, E. Zhou, and D. Metaxas, "Global matching with overlapping attention for optical flow estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 592–17 601.

[52] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *European conference on computer vision*. Springer, 2020, pp. 402–419.

[53] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.

[54] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[55] J. Huang, Z. Zhu, F. Guo, and G. Huang, "The devil is in the details: Delving into unbiased data processing for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5700–5709.

[56] Z. Geng, K. Sun, B. Xiao, Z. Zhang, and J. Wang, "Bottom-up human pose estimation via disentangled keypoint regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 676–14 686.

[57] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 568–578.

[58] W. Wang, E. Xie, X. Li, D. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "PVT v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022.

[59] Y. Cai, Z. Wang, Z. Luo, B. Yin, A. Du, H. Wang, X. Zhang, X. Zhou, E. Zhou, and J. Sun, "Learning delicate local representations for multi-person pose estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 455–472.

[60] J.-J. Liu, Q. Hou, M.-M. Cheng, C. Wang, and J. Feng, "Improving convolutional networks with self-calibrated convolutions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 096–10 105.

[61] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[62] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

Supplementary Material

In this supplementary material, we show more visualizations of anomalous skeletal data correction effect that could not be included in the main manuscript due to the lack of space.

1 Visualizations of correction effect

1.1 occlusion

We show the correction of anomalous skeletal data in the occlusion case in Fig. 1and Fig. 2. In this subsection, occlusion cases include both self-masking and external masking.

1.2 motion blur

We show the correction of anomalous skeletal data in the motion blur case in Fig. 3.

1.3 limb-joining

We show the correction of anomalous skeletal data in the limb-joining case in Fig. 3.

1.4 environmental influence

We show the correction of anomalous skeletal data in the environmental influence case in Fig. 4. Common environ- mental influences include hair, clothing, backgrounds, low lighting, etc.

1.5 common errors

We show the correction of anomalous skeletal data in the common errors case in Fig. 3 and Fig. 4. Common errors refer to false detections and missed detections.

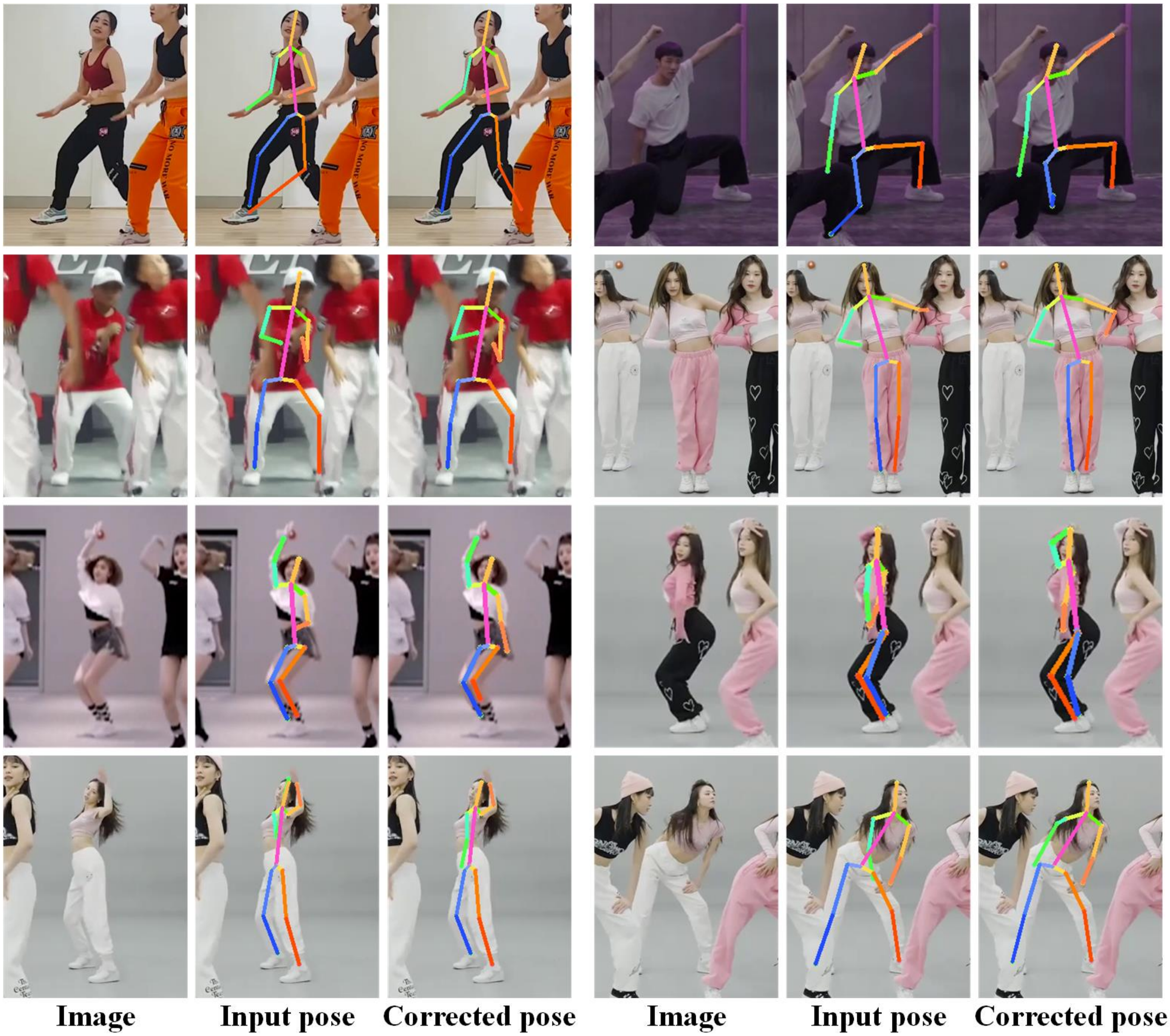


Fig. 1: Correction results of the DanceFix in the dance scene regarding the anomalies of occlusion cases. The occlusion cases include both external masking (rows 1-2) and self-masking (rows 3-4).

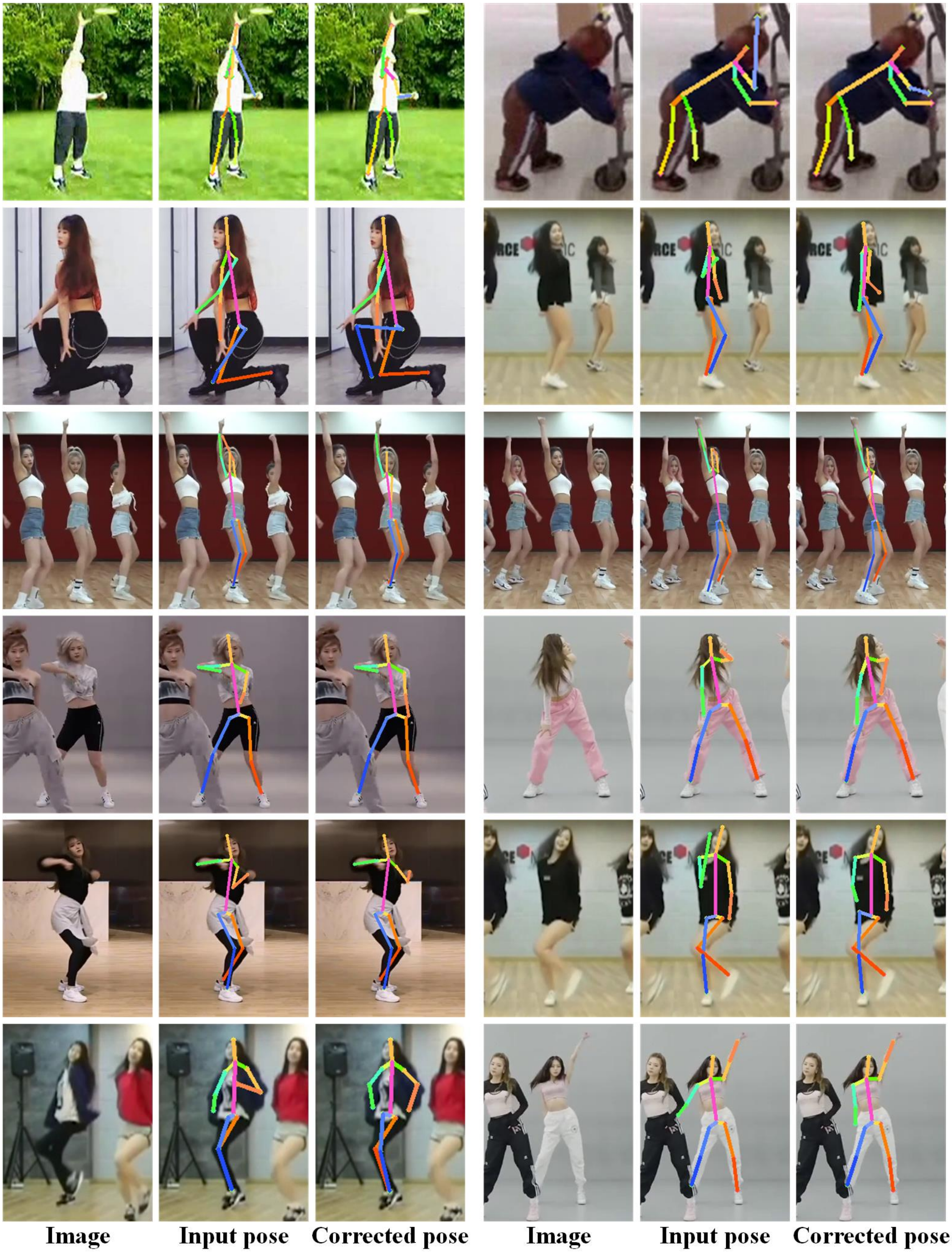


Fig. 2: Correction results of the DanceFix in the dance scene and JHMDB dataset regarding the anomalies of occlusion cases. There are all the self-masking cases.

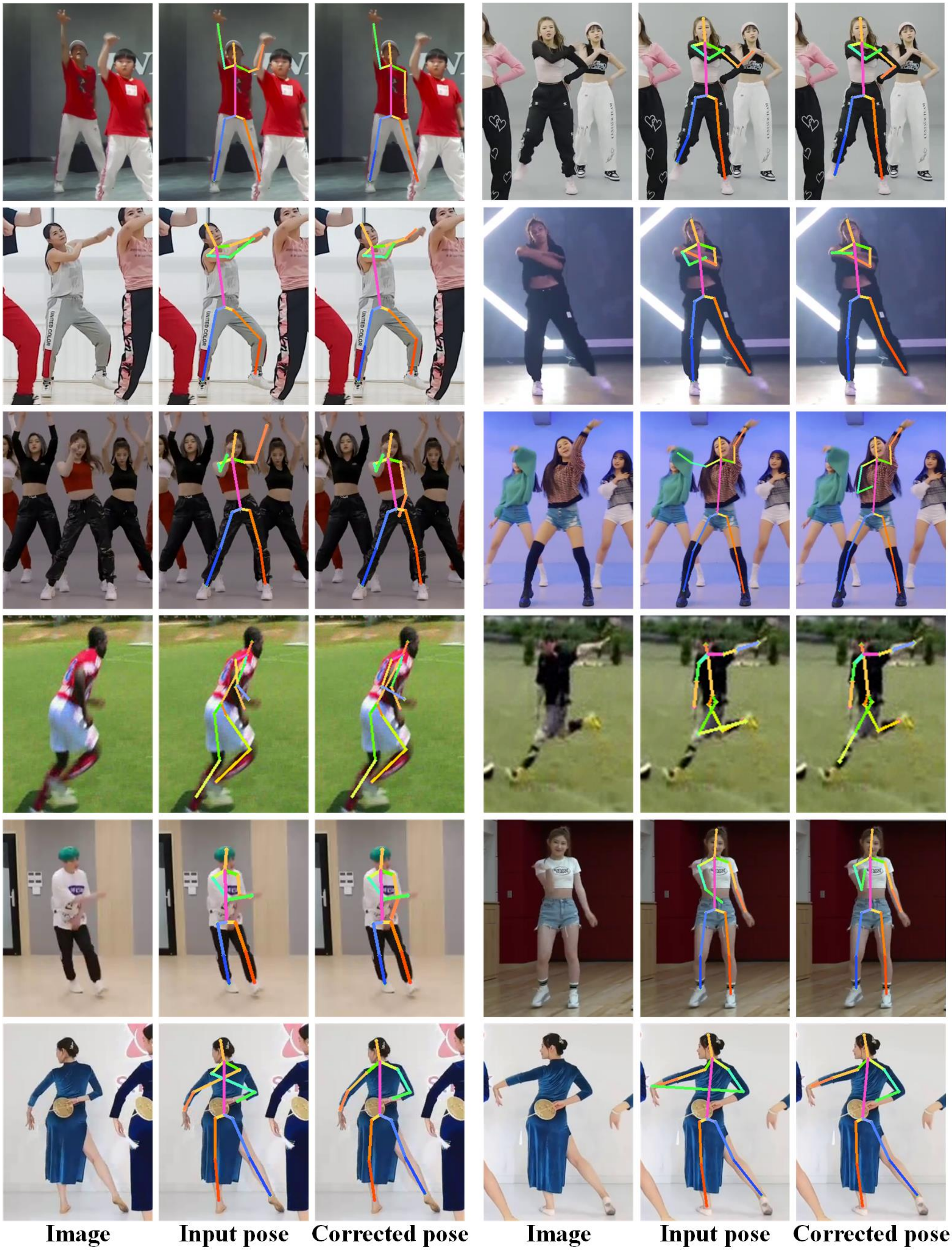


Fig. 3: Correction results of the DanceFix in the dance scene and JHMDB dataset regarding the anomalies of limb-joining (rows 1-3), motion blur (rows 4-5) and common errors (row 6).

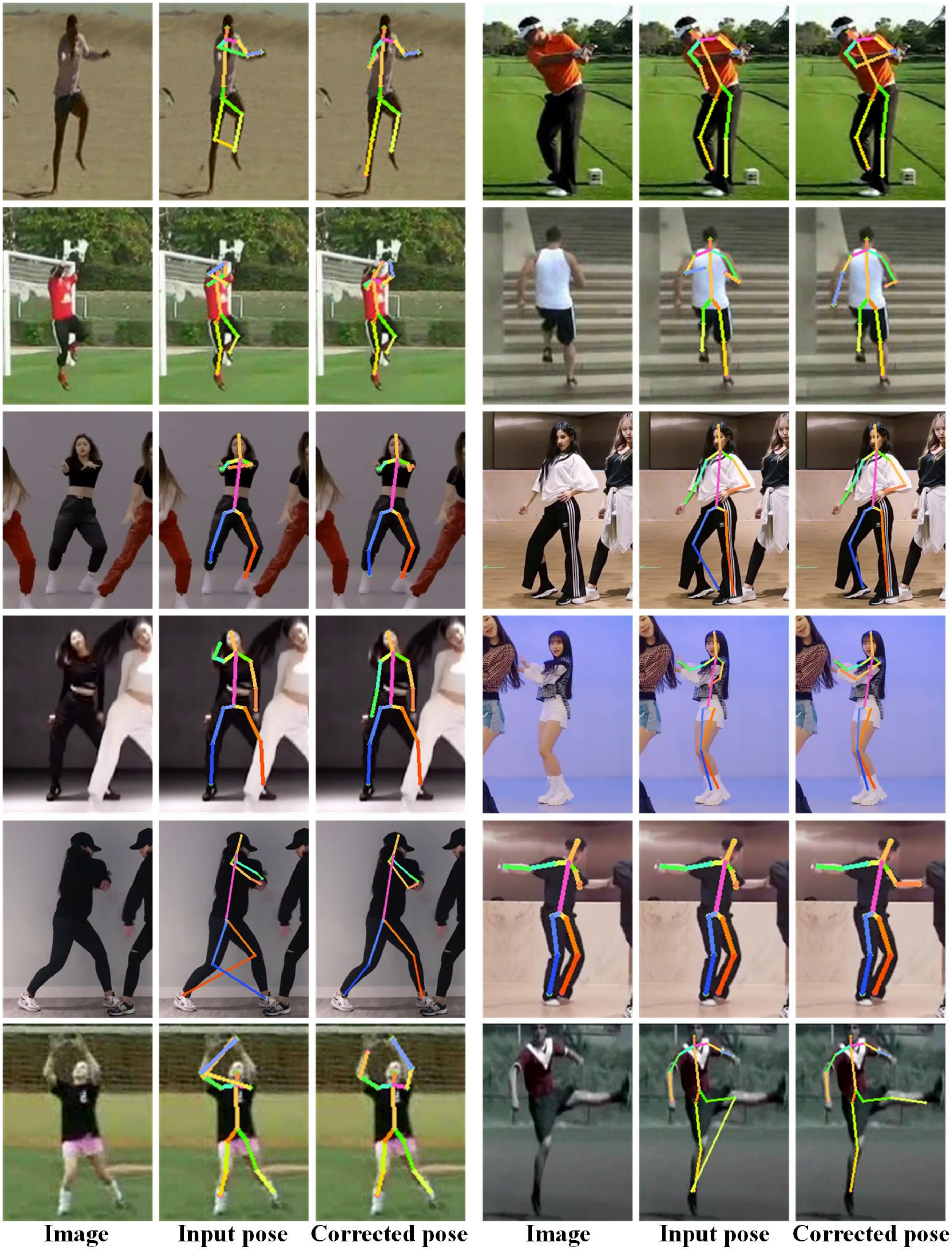


Fig. 4: Correction results of the DanceFix in the dance scene and JHMDB dataset regarding the anomalies of environmental influence (rows 1-4) and common errors (rows 5-6).